Weekly Report
09/22/2014 - 09/28/2014

Jing Xia

September 28, 2014

1 Summary

This week I mainly focused on the rank visualization project. We started the project all over again: task definition, time series distance, clustering, clustering evaluation and visualization.

2 Projects

2.1 Project 1 - Rank Visualization

2.1.1 Tasks

In the paper *Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges* [1] the authors described task related to elements, sets and element attributes of set visualization. I adapted the tasks in set visualization to our rank visualization project, and summarized 11 tasks as follows.

- Representation/Attributes of Clusters
  (1) Size
  (2) How does it form a cluster? (What is their evolving patterns in common?)
  (3) How is the cluster developed? (forming or disbanding)
- Attributes of items
  (1) What's the evolving pattern of this item (new? Up/down)?
  (2) What's the rank of this item?
- Item-Cluster Relation
  (1) Does an item belong to any cluster?
  (2) Did this item belong to any other clusters?

- Cluster-Cluster Relation
  (1) How does its pattern differ from those of other clusters?
  (2) Is it a developing cluster or a single cluster? (If a developing cluster, what's the developing pattern?)
- Rank
  (1) How's the rank evolving? Is it fluctuate or steady?
  (2) Critical time of big changes? What causes the changes?

2.1.2 Similarity and Clustering

The paper *Online Discovery of Group Level Events in Time Series* [2] discussed group level events (group formation and group disbanding) in time series with AutoDBSCAN algorithm. The DBSCAN algorithm, which extended by AuthDBSCAN algorithm in this paper, is a clustering algorithm for time series. Basically, it initializes a cluster to contain a random time series in the dataset, adding its non-classified k-distance neighbors (based on time series similarity) to the cluster, iteratively adding new neighbors of data in the cluster until there is no new time series added to the cluster. AutoDBSCAN extends the algorithm by giving a relatively large k and refining the cluster by progressively reduce the value of k for each cluster. Eventually it will get much more firmly clustered sets. This algorithm is proved to be useful in clustering spatial-related datasets to recognize the shape (stripes, spirals, etc) of the underlying data.
But I still have a concern about this disease-spreading kind of algorithm in time series clustering. Because when time series A is similar to B and B is similar to C, it does not make A similar to C. The similarity can be passed and muted in the process. I'm not sure such algorithm would be suitable for time series clustering although this paper is about time series clustering. I wrote an email to the first author for this but haven't got any reply.

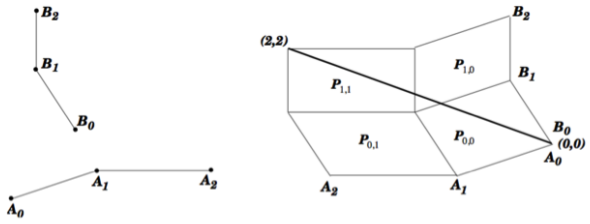

Figure 1: Polygonal chain and the corresponding manifold. The similarity of the two chains can be equivalent to distance between (0, 0) and (2, 2).

The paper *Curve Matching, Time Warping, and Light Fields: New Algorithms for Computing Similarity between Curve* [3] applies dynamic time warping in

Algorithm 1 *Compute the CDTW distance between two curves* **A** *and* **B**

Require: Curve $\mathbf{A} = \{a_1, a_2, \ldots, a_n\}$
Require: Curve $\mathbf{B} = \{b_1, b_2, \ldots, a_m\}$
Require: s: Number of Steiner points per edge
  Construct the manifold $\mathcal{M}(\mathbf{A}, \mathbf{B})$;
  Place Steiner points on $\mathcal{M}(\mathbf{A}, \mathbf{B})$;
  for $i = 2$ to n **do**
    for $j = 1$ to m **do**
      Set values for points on the bottom and right edges of the patch $\mathcal{P}_{ij}$;
      for $k = 1$ to $s + 1$ **do**
        $d(p_t^k) = \min(\arg\min_{1 \le k' \le s+1}\{d(p_b^{k'}) + \overline{|p_t^k p_b^{k'}|}\}, \arg\min_{1 \le k' \le k}\{d(p_r^{k'}) + \overline{|p_t^k p_r^{k'}|}\})$
      end for
      for $k = 1$ to $s + 1$ **do**
        $d(p_t^k) = \min(\arg\min_{1 \le k' \le s+1}\{d(p_r^{k'}) + \overline{|p_t^k p_r^{k'}|}\}, \arg\min_{1 \le k' \le k}\{d(p_b^{k'}) + \overline{|p_t^k p_b^{k'}|}\})$
      end for
    end for
  end for
  return $d(p_t^{s+1})$ computed over the patch $\mathcal{P}_{mn}$

Figure 2: The algorithm of computing the CDTW distance between two curves.

curve matching. The paper takes similarity of two polygonal chains equivalent to the optimal distance on the manifold made by the two chains (see Figure 1). The optimal distance can be calculated with dynamic time warping. In addition, to get a more accurate results, it interpolates steiner points on the edges of the manifold patches. The overall algorithm is described in Figure 2. To adapt the method we can take rank time series as polylines and compare their similarity with the method described in this paper.

2.1.3 Evaluation

The paper *Online Discovery of Group Level Events in Time Series*[2] also gave an evaluation method for quality of clustering. It is based on the entropy theory and calculates the entropy of a bunch of time series. For the entropy of a cluster χ of m time series (x_1, x_2, ... , x_m) at time t

$$S(\chi, t) = -\sum_{j=1}^{m} \frac{1}{m} \log(\sum_{i=1}^{m} \exp\left(-d(x_i, x_j, t)\right) \frac{1}{m}) \tag{1}$$

The equation 1 also takes similarity of every pair of time series in the cluster into consideration. Thus it can be a good evaluation of clustering quality.

2.2 Project 2 - Data Inspection

Not ready to summarize yet.

3 Paper Reading

Major papers have been discussed in Section 2.1, here I briefly describe some papers that I read but thought it not useful in the rank project.

Trajectory Clustering: A Partition-and-Group Framework [4] I read this paper when I first came to the idea to take time series as geometries. This trajectory clustering algorithm first partitions trajectory paths into segments based on their curvedness and then groups paths based on their segments. It is not suitable for rank time series.

Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation [5] This paper introduces an open-end dynamic time warping algorithm, which compares two time series with one as complete series and the other as partial series. This is not suitable for rank time series and cannot evaluate its effectiveness either.

Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges [1] This is a very good survey paper of set visualization. I adopted and extended its task definitions to rank visualization. None of the visualization is completely suitable for rank visualization, each has its own advantages and disadvantages. We're thinking about combining different design or coming up with new ones. Still working on it.

4 Miscellaneous

-

5 To Do List

1. Read paper [1] in detail again. Think about rank visualization design.

2. Complete implementation preparation of rank visualization.

References

[1] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges (supplementary material).

[2] Xi C Chen, Abdullah Mueen, Vijay K Narayanan, Nikos Karampatziakis, Gagan Bansal, and Vipin Kumar. Online discovery of group level events in time series. Technical report, SIAM, 2014.

[3] Alon Efrat, Quanfu Fan, and Suresh Venkatasubramanian. Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. *Journal of Mathematical Imaging and Vision*, 27(3):203–216, 2007.

[4] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.

[5] Paolo Tormene, Toni Giorgino, Silvana Quaglini, and Mario Stefanelli. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine*, 45(1):11–34, 2009.